ECE 6980
An Algorithmic and Information-Theoretic Toolbox for Massive Data

Instructor: Jayadev Acharya
Scribe: Michael Roberts

Lecture #10
27th September, 2017

1 Fingerprints

Definition 1 (Profile, aka fingerprint). *The profile of* $X_1 ... X_n$*, written* $\phi(X_1 ... X_n)$*, is a multiset* $\{N_X : X \in X_1 ... X_n\}$ *Furthermore, define* $\phi_i(X_1 ... X_n) = |\{X \in X_1 ... X_n : N_X = i\}|$

For instance, take $X_1 ... X_{11} = abracadabra$. Then $\phi(X_1 .. X_{11}) = \{1, 1, 2, 2, 5\}$. We could also specify this by noting that $\phi_1 = 2, \phi_2 = 2, \phi_5 = 1, \text{else } \phi_i = 0$

So for f symmetric, if there exists an algorithm A for estimating $f(p)$ using $X_1 ... X_n$, then there exists an algorithm A' for estimating $f(p)$ from $\phi(X_1 ... X_n)$, having the same guarantees as A.

For a fixed n, we can calculate $\mathbb{E}[\phi_i] = \sum_x p(x)^i (1 - p(x))^{n-i} \binom{n}{i}$. For, Poisson sampling, we can similarly calculate it (based on lecture 2) as $\sum_x e^{-np(x)} \frac{(np(x))^i}{i!}$

Note the polynomial terms $p(x)^i$ that appear in both of those quantities. This will motivate us to define and study a new quantity.

Definition 2. $M_\alpha(p) = \sum_x p(x)^\alpha$

We claim that for integer α these are easy to estimate, since we can do so using ϕ_α. In fact - this integer case of α is essentially all it is easy to estimate.

2 Rényi Entropy

Definition 3 (Rényi Entropy). $H_\alpha(p) = \frac{1}{1-\alpha} \log M_\alpha(p)$

Note that $\lim_{\alpha \to 1} H_\alpha = H(p)$ - just use L'Hospital's rule.

Given $X_1 ... X_n$ or $\phi(X_1 ... X_n)$, say we want to estimate $H_\alpha(p)$ up to an ϵ error. The known sample requirements for a distribution over k elements are summarized below:

- $\alpha \in \mathbb{N} : \Theta(\frac{k^{1-1/\alpha}}{\epsilon^2})$
- $\alpha \notin \mathbb{N}, \alpha \geq 1 : O(\frac{k}{\epsilon}), \Omega(k^{f(\alpha)})$ for any $f(\alpha) < 1$
- $\alpha \notin \mathbb{N}, \alpha < 1 : O((\frac{k}{\epsilon})^{1/\alpha})$

3 Approximating Entropy

Take any degree-d polynomial $P(y) = \sum_{i=0}^{d} a_i y^i$. Then we can approximate $\sum_x P(p(x)) = \sum_x \sum_{i=0}^{d} a_i p(x)^i = \sum_{i=0}^{d} a_i M_i(p)$, since we can approximate each $M_i(p)$. It's too bad that entropy is not a polynomial!

Luckily, the problem of approximating arbitrary functions by polynomials of bounded degree is well-studied. The area is called approximation theory.

Definition 4. *The best polynomial approximation of f on $[a, b]$, of degree d, is*
$\arg\min_{P \in P_d} \sup_{x \in [a,b]} |f(x) - P(x)|$
where P_d is the set of real polynomials of degree d.

So if we take P to be the best polynomial approximation of degree d of the function $f(y) = y \log \frac{1}{y}$, then we can approximate $H = \sum_x f(p(x))$ by $\sum_x P(p(x))$, and approximate that using the above strategy.

Note that we haven't specified what degree d we will use. This will have to be chosen with some care. On one hand, increasing d will mean that our polynomial approximation will have lower error. On the other, since the sample complexity of M_α for integer α increases with α, increasing d will require us to take more samples.